# Inference with few assumptions: Wasserman's example

Christopher A. Sims
Princeton University
sims@princeton.edu

October 27, 2007

**Types of "assumption-free" inference**

- A simple procedure or set of statistics is proposed. It is shown to have "good properties" under "very weak" assumptions about the model.
  - OLS, IV, GMM, differencing out individual effects
- An infinite-dimensional parameter space or a nuisance-parameter vector that grows with sample size is explicit. A procedure that explicitly accounts for the infinite-dimensionality is proposed.
  - Kernel regression, frequency-domain time series models.

**Are these approaches inherently non-Bayesian?**

- Certainly the idea that we would prefer that inference not depend on uncertain or arbitrary assumptions is reasonable from any perspective.
- Some Bayesians have argued that insistence on using all aspects of the observed data is an important part of the Bayesian perspective on inference, but I don't agree.
  - Simple procedures that "work" will be used. Bayesians might want to use them for the same reason non-Bayesians might.

– Spuriously appealing simple procedures should not be used. Bayesian analysis of simple procedures may help us identify procedures, or combinations of procedures and types of samples, in which the simple procedures break down.

**What are micro-founded (i.e. Bayesian) versions of these approaches?**

- Limited-information Bayesian inference based on asymptotics. (Yum-Keung Kwan). Convert frequentist asymptotics to posteriors conditional on the statistics used to form the estimator and its asymptotic variance.
- Bayesian sieves. Postulate a specific prior distribution over an infinite-dimensional parameter space and proceed. Usually this has the form of a countable union of finite-dimensional parameter spaces with probabilities over the dimension.
- Limited-information Bayesian inference based on entropy. "Conservative" inference based on the idea that assumptions that imply the data reduce entropy slowly are weaker than assumptions that imply that in this sense information accumulates faster. (Jaynes, Zellner, Jae-young Kim)

**Explicit infinite-dimensionality**

- Uncertain lag length; polynomial regression with degree uncertain; autoregressive estimation of spectral density with uncertain lag length; random effects approaches to panel data.
- These appear to be restrictive. Insisting that a regression function is a finite-order polynomial with probability one seems more restrictive than just insisting that, e.g., there is some (unknown) uniform bound on some of its derivatives.
- That's not true, though. *Every* Borel measure on a complete, separable metric space puts probability one on a countable union of compact sets, and in an infinite-dimensional linear space, compact sets are all topologically small in the same sense that finite-dimensional subspaces are.
- Frequentist procedures that provide confidence sets face the same need to restrict attention to small subspaces.

- Example: $\ell_2$, the space of square summable sequences $\{b_i\}$. Many time series setups (most frequency domain procedures, e.g.) require that there is some $A > 0$ (which may not be known) such that $A\,|b_i|\,i^q$ is bounded for some particular $q \geq 1$. The set of all such $b$'s is a countable union of compact sets. Compact subsets of an infinite dimensional linear space are topologically equivalent to subsets of $\mathbb{R}^k$. So the union over $T$ of all $b$'s such that $b_i = 0$ for $i > T$ is the same "size" as the subsets satisfying this rate of convergence criterion.

- Time series econometricians long ago got over the idea that frequency domain estimation, which makes only smoothness assumptions on spectral densities, is any more general than time-domain estimation with finite parametrization. In fact, the preferred way to estimate a spectral density is usually to fit an AR or ARMA model.

- It's a bit of a mystery why non-time-series econometricians still habitually suggest that kernel estimates make fewer assumptions than (adaptive) parametric models.

**No Lebesgue measure**

- $\mathbb{R}^n$ is a complete, separable metric space. But Lebesgue measure provides a topologically-grounded definition of a big set as an alternative to "category". Lebesgue measure is Borel and translation-invariant.

- In an $\infty$-dim space $S$, for any Borel measure $\mu$ on $S$, there is an $x \in S$ such that for some measurable $A \subset S$, $\mu(A) > 0$ and $\mu(A + x) = 0$. I.e., not only is there no translation-invariant measure, translation can always take a measure into one that is mutually discontinuous with it.

**Is this a problem?**

- Not a special problem for Bayesians. Any attempt to produce confidence regions runs into it.

- We can never relax about the possibility that a reasonable-looking prior is actually dogmatic about something important. In $\mathbb{R}^n$, priors equivalent to Lebesgue measure all agree on the zero-probability sets and in this sense are not dogmatic. No such safe form of distribution is available in $\infty$-dimensions.

**Using entropy**

- While entropy-reduction has axiomatic foundations, using it as if it came from a loss function is hazardous. See Bernardo and Smith.
- We may care about some parameters and not others, e.g., so that assumptions that let us learn rapidly about parameters we don't care about, while limiting the rate at which we learn about others, may in fact be conservative.
- One reasonable approach: Make assumptions about model *and prior* that imply a joint distribution for the observations and the data that makes the **mutual information** between data and parameters as small as possible, subject to some constraints.

**Minimizing mutual information between $y$ and $\beta$ s.t. a fixed marginal on $y$**

- This is often easy, and appealing.
- When maximizing the entropy of $y \mid \beta$ leads to a well-defined pdf, this solves the problem.
- E.g. the SNLM, which emerges as maximum entropy given the conditional mean and variance of $y$
- The fixed marginal on $y$ is perhaps fairly often reasonable — if we know nothing about parameters, we often know almost nothing about the distribution of $y$.
- IV: emerges from the IV moment conditions. Leads to using the LIML likelihood to characterize uncertainty, which makes a lot more sense than treating the conventional asymptotic distribution for IV itself as if it were correct.
- This approach handles weak instruments automatically.

**The Wasserman example**

- Why study this very stylized example?
- It originally was meant to show that Bayesian or other likelihood-based methods run astray in high-dimensional parameter spaces, while a simple, frequentist, GMM-family estimate worked fine — something that some statisticians and econometricians think is well known, which is probably why Wasserman, a very smart statistician at CMU, himself ran astray in the example.

- The model and estimator are a boiled-down version of "propensity score" methods that are widely used in empirical labor economics.
- The discussion will illustrate all the general points made above.
- For each $\omega \in \{1, \ldots, B\}$ there is a pair $\theta_i, \xi_i$, each of which is in $[0, 1]$. $B$ is a very large integer.
- We know $\xi() : \{1, \ldots, B\} \mapsto (0, 1)$.
- We do not know $\theta() : \{1, \ldots, B\} \mapsto (0, 1)$, though it exists. Our inference will concern this unknown object.

**The sample**

- For each observation number $j = 1, \ldots, N$, an $\omega_j$ was drawn randomly from a uniform distribution over $\{1, \ldots, B\}$.
- $R_j \in \{0, 1\}$, $P[R = 1 \mid \xi(\omega_j), \theta(\omega_j)] = \xi(\omega_j)$.
- (From this point on we will use the shorthand notation $\theta_j$ and $\xi_j$ for $\theta(\omega_j)$ and $\xi(\omega_j)$).
- $Y_j \in \{0, 1\}$, $P[Y_j = 1 \mid \theta_j, \xi_j] = \theta_j$
- We observe the triple $(\xi_j, R_j, R_j Y_j)$.
- We are interested in inference about $\bar{\theta} = (1/B) \sum_\omega \theta(\omega)$

**The setup in words**

We have a sample of a zero-1 variable $(Y_j)$ with missing observations. We would like to know the average probability that $Y$ is one, but we can't simply take sample averages of the $Y_j$'s, because the probability that a sample value $Y_j$ is missing might vary with $\theta_j$.

**Wasserman's claims**

The pdf of our observations as a function of the unknown parameters $\{\theta(\omega), \omega \in 1, \ldots, B\}$ (and of the known parameters $\xi(\omega)$) is

$$\prod_{j=1}^{N} \xi_j^{R_j} (1 - \xi_j)^{1 - R_j} \theta_j^{R_j Y_j} (1 - \theta_j)^{(R_j - R_j Y_j)} .$$

The likelihood function, which varies only with the unknown parameters $\theta(1, \ldots, B)$, depends on the known $\{\xi_j\}$ sequence only through a scale factor, which does not affect the likelihood shape. Therefore Bayesian, or any likelihood-based inference, must ignore the $\xi_j$'s. If $N << B$, most of the $\theta(\omega)$ values do not appear in the likelihood function. For all those values of $\omega$ that have not appeared in the sample, the posterior mean has to be the prior mean. The posterior mean for $\bar{\theta}$ must therefore be almost the same as the prior mean for that quantity.

**The robust frequentist approach**

- Use the **Horwitz-Thompson** estimator.
- This just weights all observations (including those for which $Y$ is unobserved, treating them as zeros) by the inverse of their probability of inclusion in the sample. That is, set
$$\hat{\theta} = \frac{1}{N} \sum_{j=1}^{N} \frac{R_j Y_j}{\xi_j}.$$
- It's easily checked that if $Z_j = R_j Y_j / \xi_j$, $E[Z_j] = E[\theta_j]$. So $\hat{\theta}$ is unbiased and strongly consistent.

**"Conservative" frequentist confidence intervals**

- If we have a lower bound on $\xi$, we have an upper bound on the variance of $\hat{\theta}$, which will allow us to create Chebyshev-style conservative confidence intervals for $\bar{\theta}$, valid even in small samples.
- Obviously we can also create the usual sort of "asymptotically valid" confidence interval (which may not have approximately correct coverage at any sample size) by forming the usual sort of $t$-statistic from the $Z_j$'s,

**What about the likelihood principle? The wrong likelihood.**

- Doesn't Wasserman's argument show that inference about $\bar{\theta}$ that uses the $\xi_j$ values violates the likelihood principle?
- Wasserman has actually incorrectly specified the likelihood function. What he (and we, above) displayed is the correct conditional joint pdf for $(Y_j, R_j Y_j)$ |

$\{\xi_j, \theta_j,\, j = 1, \ldots, N\}$. It is also true that $\{\theta(\omega), \xi(\omega), \omega = 1, \ldots, B\}$ can be regarded as unknown and known "parameters", respectively. But the $\xi_j$'s are not the $\omega_j$'s. The $\xi_j$ are realized values of a random variable that is observed. The likelihood must describe the joint distribution of all observable random variables, conditional on the unknown parameter values, which here are just $\{\theta(\omega)\}$.

**Correcting the likelihood**

- So we need to specify the joint distribution of $\theta_j, \xi_j$, not just condition on them, and then, since the $\theta_j$'s are unobservable random variables, we have to integrate them out to obtain the conditional distribution of observables given the "parameters" $\{\theta(\omega)\}$.
- The probability space is $S = \{1, \ldots, B\}$. The joint distribution is fully characterized by $\xi(\omega)$ and $\theta(\omega)$. But since the $\theta(\omega)$ function is unknown, we can't write down the pdf of the data, and therefore the likelihood, without filling in this gap.

**Ways to fill the gap**

- Postulate independence. Then the likelihood, and hence inference about $\theta()$ will not depend on the observed values of $\xi_j$, indeed inference can be based on the sample where $Y_j$ is observed, ignoring the censoring.
- However even in this case Wasserman is wrong to say that the form of the likelihood implies that the posterior mean for $\theta(\omega)$ must be the prior mean for all those values of $\theta(\omega)$ that do not appear in the sample.

- The prior on $\theta()$ in the independence (of $\xi$) case is a stochastic process on $1, \ldots, B$. It can perfectly well be an *exchangeable* process with uncertain mean. For example,

$$\{\theta(\omega) \mid \nu\} \sim \text{Beta}(k\nu, k(1-\nu)),\ \text{all } \omega$$
$$\nu \sim U(0,1)\,.$$

- With this prior inference is straightforward and the posterior mean of $\bar{\theta}$ will be very close to the average value of $Y_j$ over those observations for which it is observed.

- But notice: We at first inadvertently, by extending a reasonable low-dimensional prior to a high-dimensional space, were dogmatic about exactly what interested us. We fixed that.
- But we're still dogmatic about a lot of other stuff — e.g., the difference in the average of the $\theta$'s over the first half and the last half of the list of $\theta$'s.
- We should probably be modeling our prior on $\theta(\omega)$ as a stochastic process.
- But it is inevitable that we are in some dimensions dogmatic.

**Limited information**

- Horwitz-Thompson achieves simplicity — both for calculating the estimator and for proving a couple of its properties (unbiasedness and consistency) by throwing away information.
- Throwing away information to achieve simplicity is sometimes a good idea, and is as usable for Bayesian as for non-Bayesian inference.
- What would Bayesian inference based on $Z_j = R_j Y_j / \xi_j$, $j = 1, \ldots, N$ alone look like?
- $Z_j$ has probability mass concentrated on a known set of points: $\{0, 1/\xi(\omega), \omega = 1, \ldots, B\}$. We don't know the probabilities of those points for a given draw $j$ conditional on $\{\theta()\}$.

**Small number of $\xi(\omega)$ values**

- If the range of the $\xi(\omega)$ function is only a few discrete points, within the set of $\omega$ values for which $\xi_j$ takes on a particular value, there is trivially no dependence of $\theta_j$ on $\xi_j$.
- It therefore makes sense to apply our solution above for the independence case separately to the $k$ subsamples corresponding to the $k$ values of $\xi_j$. The estimated mean $\theta$'s within each subsample can then be weighted together using the assumed-known probabilities of the $k$ $\xi(\omega)$ values.

**Extension to many $\xi(\omega)$ values**

- We could break up the range of $\xi(\omega)$ into $k$ segments, assume the distribution of $\theta(\omega)$ within each such segment is independent of $\xi(\omega)$. This obviously is feasible and will be pretty accurate if the dependence of the distribution of $\theta \mid \xi$ on $\xi$ is smooth.
- To be more rigorous, we could put probabilities $\pi_k$ over the number of segments, with each $k$ corresponding to a given list of $k$ segments of (0,1).
- This is a Bayesian sieve, and leads to accurate inference under fairly general assumptions.
- This approach would work even if there is an uncountable probability space and we are just given a marginal density $g(\xi)$ for $\xi$.

**Using entropy**

- It makes things easier here if we (realistically) suppose that the underlying probability space is uncountable, rather than $\{1, \ldots, B\}$.
- Each $Z_j$ has an unknown distribution on $\{0\} \cup [1, \infty)$, but with mean $\bar{\theta}$.
- We can often arrive at conservative Bayesian inference by maximizing entropy of the data conditional on known moments. Here, the max entropy distribution for $Z$ will put probability $\pi_0(\bar{\theta})$ on $Z = 0$ and density $\pi_0(\bar{\theta}) \exp(-\lambda(\bar{\theta})Z)$ on $Z \geq 1$.

- The max entropy distribution makes $\sum Z_j$ a sufficient statistic. Its posterior mean will be consistent and respects the a priori known bound $\theta \in [0, 1]$.
- If we relax the constraint connecting $\pi_0$ and the height of the continuous pdf at $Z = 1$, we get a sample likelihood

$$\alpha^n (1-\alpha)^m \mu^n e^{-\mu \sum_{Z_i > 0} (Z_i - 1)} ,$$

  where $n$ is the number of non-zero observations and $m$ the number of zero observations on $Z$.
- The bounds on $\theta$ place bounds on $\mu, \alpha$.

- Ignoring the bounds and using a conjugate prior parametrized by $\gamma$ gives a posterior mean for $\bar{\theta}$ of
$$\hat{\theta} = \frac{n+1}{n+m+2} \cdot \left( \frac{\sum Z_i}{n} - \frac{\gamma}{n} \right) ,$$
very close to the Horwitz-Thompson estimator for large $n, m$.

**Admissibility**

- Ignoring the bounds makes this estimator — and the Horwitz-Thompson estimator — inadmissible. If $\theta$ is near 1, these estimators can have a high probability of exceeding 1.
- We can truncate them, but this undoes unbiasedness of Horwitz-Thompson.
- One admissible estimator is obtained as the posterior mean of the $\alpha, \mu$ model, accounting for the bounds.
- The draft note shows how that model leads to a posterior that is more diffuse when there are fewer non-zero $Z$'s.
- The entropy-maximizing model would not have this property. Its posterior would depend on $\sum Z_j$ and sample size, but not on the number of non-zero draws.
- Probably, therefore, its posterior is more diffuse for samples with few-non-zero draws. I haven't decided yet whether this is going to often be reasonable.

**Lessons**

I. Don't try to prove that the likelihood principle leads to bad estimators.

II. In high-dimensional parameter spaces (as with nuisance parameters), be careful to assess what kind of prior knowledge is implied by an apparently standard-looking prior. Use exchangeability to make sure important functions of the distribution are reasonably uncertain a priori.

III. Good estimators usually are exactly or almost equivalent to Bayesian estimators. Interpreting them from a Bayesian perspective can often suggest ways to improve them.

**GMM**

- If one minimizes mutual information between $y$ and $\beta$ subject to $E[g(y;\beta)] = 0$, $E[g(y;\beta)g(y;\beta)' = \Sigma]$ and to a fixed marginal on $y$, one arrives at a pdf for $\{y \mid \beta, \Sigma\}$ of the form

$$\exp(-\lambda(y) - \mu(\beta,\Sigma)g(y;\beta) - g(y;\beta)'\Omega(\beta,\Sigma)g(y;\beta)) \,,$$

  where $\lambda$ and $\mu$ have to be chosen to meet the requirement that the pdf integrates to one for all $\beta, \Sigma$ and the moment condition is satisfied.

- Note this does not mean $g$ is itself Gaussian, and $\Omega \neq \Sigma$. (except in linear models).

- Looking for $\lambda$, $\mu$ and $\Omega$ for some interesting real-world examples of nonlinear $g$'s is an interesting project.